Estimating the control parameters of an articulatory model from electromagnetic articulograph data

Asterios Toutios,[a] Slim Ouni, and Yves Laprie
Laboratoire Lorrain de Recherche en Informatique et ses Applications, Unité de Recherche Mixte 7503, Boîte Postale 239, 54506 Vandœuvre-lès-Nancy Cedex, France

(Received 17 June 2009; revised 24 February 2011; accepted 25 February 2011)

Finding the control parameters of an articulatory model that result in given acoustics is an important problem in speech research. However, one should also be able to derive the same parameters from measured articulatory data. In this paper, a method to estimate the control parameters of the the model by Maeda from electromagnetic articulography (EMA) data, which allows the derivation of full sagittal vocal tract slices from sparse flesh-point information, is presented. First, the articulatory grid system involved in the model's definition is adapted to the speaker involved in the experiment, and EMA data are registered to it automatically. Then, articulatory variables that correspond to measurements defined by Maeda on the grid are extracted. An initial solution for the articulatory control parameters is found by a least-squares method, under constraints ensuring vocal tract shape naturalness. Dynamic smoothness of the parameter trajectories is then imposed by a variational regularization method. Generated vocal tract slices for vowels are compared with slices appearing in magnetic resonance images of the same speaker or found in the literature. Formants synthesized on the basis of these generated slices are adequately close to those tracked in real speech recorded concurrently with EMA. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3569714]

PACS number(s): 43.70.Bk, 43.70.Aj, 43.70.Jt [SSN] Pages: 3245–3257

I. INTRODUCTION

Articulatory models[1–5] describe the vocal tract shape by means of a small number of control parameters. Given these parameters, the models can produce vocal tract area functions and derive corresponding acoustic outputs. However, the values of the parameters that produce a desired acoustic target or a desired vocal tract shape are usually not known beforehand. Obviously, the exceptions are the values corresponding to the coupled articulatory and acoustic data that were used for the construction of the model. In other words, the exact way a given articulatory model should be driven in order to produce a desired output is often unknown.

Nevertheless, knowing how to drive an articulatory model is essential, especially toward the goal of articulatory speech synthesis which holds the promise of improving the naturalness of computer generated speech.[6] In many ways, without the knowledge of how to drive it, an articulatory model is more or less like a musical instrument without a musical score.

Model-based acoustic-to-articulatory inversion[6] aims at recovering articulatory control parameters given acoustics. The most common paradigm is the analysis-by-synthesis approach where some optimization technique, often exploiting a codebook,[7,8] is used to minimize the acoustic distance between the target acoustics and those produced by an articulatory synthesizer. The articulatory synthesizer is a realization of an articulatory model, and the optimization process leads to values for its control parameters. Such methods are commonly evaluated only in the acoustic space, after resynthesis from the recovered articulatory representation. However, derived articulatory control parameters should also agree with measured articulatory data.

Ideally, the articulatory information to be used for such a purpose should be related to continuous contours of the full midsagittal vocal tract shape. Today, such information is only available using magnetic resonance imaging (MRI) (Refs. 9 and 10) since x-ray imaging has been abandoned in the eighties due to the health hazard linked to exposure of human subjects to radiation,[11] and because ultrasound-based techniques provide only partial tongue contours.[12] However, despite its acknowledged innocuousness, MRI suffers currently from very slow acquisition rates, which makes it useful only for sustained articulations. Moreover, the subject has to lay supine during MRI acquisition; this induces certain articulatory artifacts compared to natural (upright) articulation.[13] Finally, noise in the scanner can also be a problem, since the subject does not have the appropriate auditory feedback.

An articulatory acquisition technique that allows more natural speech production and offers high acquisition rates is electromagnetic articulography (EMA).[14–17] Nevertheless, it concerns only the movement of several sensor coils attached to the articulators. The number of sensor coils that can be glued on the tongue is limited to three or four, for minimal interference with natural articulation and minimal electromagnetic interactions between coils. Furthermore, it is very difficult to place sensor coils in the area of the tongue root. These limitations make it hard to obtain directly the whole midsagittal configuration of the tongue.

[a]Author to whom correspondence should be addressed. Electronic mail: asterios.toutios@loria.fr

The goal of our work is to derive articulatory control parameters from EMA data, which is in many aspects equivalent to deriving full continuous midsagittal vocal tract contours. Kaburagi and Honda[18] first showed that regions of the tongue contour between EMA sensors can be efficiently predicted from EMA data. They described a method which mapped successfully EMA information for a Japanese speaker onto continuous ultrasound tongue contours, which were synchronously recorded and registered.

Several works have extended this idea to examining the efficiency of inferring midsagittal pharyngeal shape from measurements on the anterior part of the tongue. Badin *et al.*[19] built an articulatory model from x-ray films of a male French speaker using an articulatory grid that adjusted dynamically its coordinates in order to follow the movements of the larynx and the tongue tip. Then, they inverted that model in order to predict full sagittal slices on the basis of their intersections with chosen combinations of three gridlines. Their main goal was to determine the optimal positions to glue EMA sensor coils in a future experiment. Whalen *et al.*[20] performed regression analysis to assess whether pharyngeal widths, as measured on MRI images of two American speakers, could be predicted from either the locations and measurements of four tongue fleshpoints or from categorical phonetic features describing tongue position and height, or from a combination of both. They reported high overall predictability and suggested that a small number of measurements on the tongue could be sufficient to guide the modeling of the pharynx for articulatory synthesis of speech. Jackson and McGowan[21] described a similar approach using x-ray data for four Swedish speakers. Their main contribution was to show that the anterior tongue shape information could be reduced to three factors and still be able to predict pharynx dimensions. These works did not use actual EMA sensors but rather marks on the x-ray or MRI images (Badin *et al.*[19] referred to these marks as *synthetic pellets*).

The present study differs from the aforementioned ones in several aspects. We do not create a new model for our speaker's articulation but use the well established model by Maeda,[3,22] adapted to our speaker. The adaptation consists of determining the mouth and pharynx geometrical scale parameters and is based on a single MRI image of the speaker. We use actual recorded EMA information and not marked flesh-points. Beside tongue contours, we are equally interested in the dynamics of the articulatory control parameters of the model by Maeda. We thus estimate control parameters that govern the positioning of the lips, in addition to the ones that govern the tongue contour. Additionally, we experiment in synthesizing speech from these parameters.

The model by Maeda is widely accepted by the speech community and has been used in several studies. However, it has several limitations. First, it was built using data from a certain female speaker of French. Though it can be adapted easily to new speakers by modifying pharynx and mouth sizes, its relevance is not ensured for sounds not present in the French language. It can also present problems if the articulatory strategy of the speaker under study is considerably different from the strategy of the speaker the model was built on due to anatomical differences or just different articulatory preferences. Moreover, the performance of the model for consonantal sounds is not considered very good, and there is ongoing work to improve it. Given that the model describes the tongue using only three parameters, there is not enough flexibility to represent accurately the anterior part of the tongue for dental, alveolar, and postalveolar consonants. Some gestures, like retroflex ones, were not covered by the corpus used by Maeda.

While we acknowledge such limitations, we consider the model by Maeda as a reasonably good model for French vowels, and our study in this paper is restricted to them. We also make the working assumption that, after determining mouth and pharynx scale factors and correcting the external tract contour, it is relevant for our subject, a male speaker of French with a different tract size and palatal shape. We believe that, even under such restrictions and assumptions, it is important to study the behavior of the model (as any articulatory model) in the face of articulatory data different than the data used to build it.

In what follows, we first revisit the model by Maeda to outline a series of aspects that are relevant to our method. We then present in detail our method for recovering the articulatory parameters of Maeda from EMA data, including information on our EMA acquisition setup. We continue by presenting examples of derived articulatory control parameter trajectories, vocal tract shapes, and synthesized speech. Derived vocal tract shapes for vowels are compared to MRI images of the same speaker, as well as to vocal tract shapes found in the literature, and synthesized formants are compared to formants tracked in real speech that was recorded concurrently with EMA. Findings from these comparisons are discussed in the context of evaluating our method.

II. PRESENTATION OF THE ARTICULATORY MODEL

The model by Maeda[3,22] describes the midsagittal slice of the vocal tract in the form of a weighted sum of seven linear components [see Fig. 1(a)]. Each weight, i.e., each articulatory parameter, is centered and normalized by dividing it by its standard deviation, thus varying roughly between -3 and 3.

One linear component gives the jaw position, three describe the tongue (tongue dorsum position, tongue dorsum shape, and apex position), two describe the lips (opening and protrusion), and the last component corresponds to the larynx height. The components were derived by applying factor analysis to articulatory contours. These contours were extracted by hand from x-ray images of vowels and parameterized by projection onto a special semipolar coordinate grid system.

The semipolar coordinate grid system consists of three regions [see Fig. 1(b)]: a linear region in the buccal area, a polar region in the velar area, and a second linear region in the pharyngeal area. In the polar region, the coordinate grids are spaced by 11.25°. The spacing of the grids in the two linear regions depends on the size of the vocal tract of the

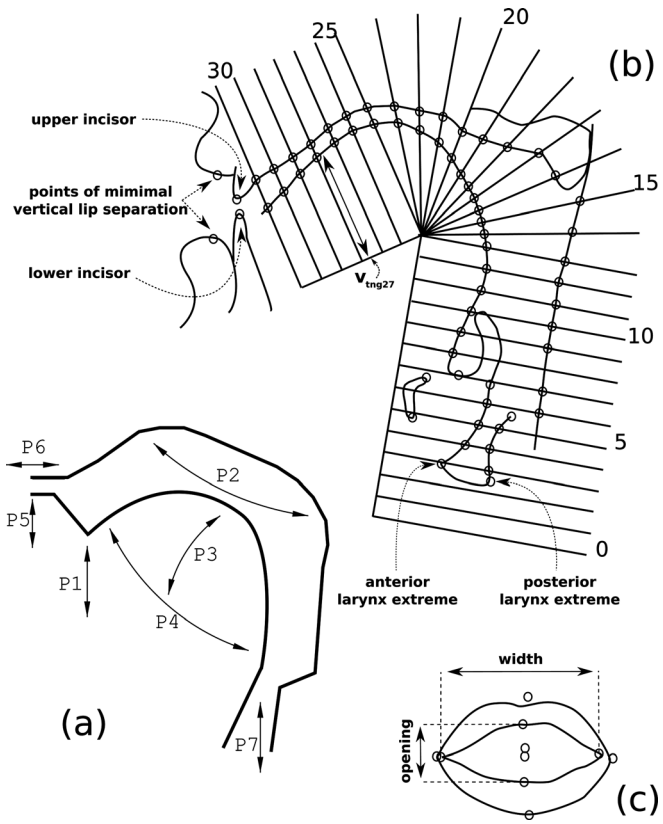

FIG. 1. (a) Parameters of articulatory model by Maeda: P1 jaw position, vertical movement, P2 tongue dorsum position that can move roughly horizontally from the front to the back of the mouth cavity, P3 tongue dorsum shape, i.e., rounded or unrounded, P4 apex position; this parameter deforms the apex part of the tongue (by moving it up or down) as well as the root of the tongue (anterior and posterior movement, respectively), P5 lip opening, P6 lip protrusion, P7 larynx height. (b) Vocal tract profile superimposed with a semipolar coordinate system, where the circles indicate the measured points. (c) Frontal view of the lips. Figure partly reproduced from the work by Maeda (Ref. 22). Several details are annotated. See text (Sec. II) for more explanations.

speaker in question. Two scale factors are enough to describe these morphological characteristics: the *mouth scale factor* and the *pharynx scale factor*.

The grid system is fixed to the rigid maxillary structure. The interior contour consists of the tongue tip, body, root, and the upper part of the larynx. The exterior wall consists of the upper incisors, hard and soft palates, and pharyngeal walls. The exterior wall is considered rigid. The grid system is supplemented with a schema of the frontal view of the lips [Fig. 1(c)].

Measurements on the grid system and the frontal view of the lips are collectively called *variables*. The *tongue variables* are defined as the coordinates of the intersections of the tongue contour with the grids. Out of the 31 gridlines, numbered as shown in Fig. 1(b), the 6th to 30th are related to the tongue contour. Thus, variables v_{tng6}, …, v_{tng30} are defined. In Fig. 1(b), the measurement of v_{tng27} is shown as an example. The 0th to 5th gridlines correspond to the area where the extremes of the larynx move. The variables describing the form of the larynx are the (x,y) coordinates of the anterior $(v_{a,x},v_{a,y})$ and posterior $(v_{p,x},v_{p,y})$ extremes of the larynx with respect to the linear coordinate system. The *lip opening variable* (v_{ope}) is defined as the distance between the highest and lowest points on the front inner lip contours. The *lip width variable* (v_{wid}) is defined as the distance between the most left and right points on the same contours. The *lip protrusion variable* (v_{pro}) is measured on the lip profile as the distance between the upper incisors and the point of the minimal vertical separation between the upper and lower lips. The *jaw variable* (v_{jaw}) is defined as the negative of the distance between the upper and lower incisors, projected on the direction of the lines of the linear region of the grid in the buccal area. All variables are normalized (z-scored).

According to the definition of the articulatory model, the variables described above are generated, at any given instant, from an underlying set of model parameters via a set of linear relationships. More specifically, the model provides the matrices A_{tng}, A_{lip}, and A_{lrx}, so that

$$\begin{aligned}
\left[v_{jaw}, v_{tng6}, v_{tng7}, \ldots, v_{tng30}\right]^T &= \mathbf{A}_{tng}\left[P_1, P_2, P_3, P_4\right]^T \\
\left[v_{jaw}, v_{pro}, v_{ope}, v_{wid}\right]^T &= \mathbf{A}_{lip}\left[P_1, P_5, P_6\right]^T, \\
\left[v_{jaw}, v_{a,x}, v_{a,y}, v_{p,x}, v_{p,y}\right]^T &= \mathbf{A}_{lrx}\left[P_1, P_7\right]^T,
\end{aligned} \tag{1}$$

where P_1 is the jaw position parameter (i.e., the weight that corresponds to the linear component describing jaw position); P_2 is the tongue dorsum position parameter; P_3 is the parameter describing tongue dorsum shape; P_4 is the tongue apex parameter; P_5 is the lip opening parameter; P_6 is the lip protrusion parameter; and P_7 is the larynx height parameter.

III. DERIVING MODEL PARAMETERS FROM EMA

A. Acquisition of EMA data

We recorded EMA data using the AG500 articulograph.[16] This apparatus provides the three-dimensional (3D) positions, azimuth, and elevation of 12 rectangular sensor coils, sampled at 200 Hz. The coordinate reference system for these measurements is a fixed cube box wherein the head of the speaker may move freely. Three reference sensors were used for the compensation of this movement. They were glued on the bridge of the nose (between the eyes) and behind the ears. Four sensors were glued on the tongue, on a line roughly on the midsagittal plane: one sensor coil approximately on the tongue tip and three more at 1.4, 3.1, and 5.7 cm from it toward the tongue root. We have observed experimentally that the results of the method presented here are best when at least one sensor is glued sufficiently toward the back of the tongue, well inside the polar region of the grid system by Maeda. Two sensors were placed on the two lip corners at the junctions between the upper and lower lip, one sensor on the middle of the upper lip, one on the middle of the lower lip, and one on the lower incisor (see Fig. 2).

The recorded data included several vowel–vowel and vowel–consonant–vowel sequences, as well as a set of 200 phonetically balanced short French sentences.[23]

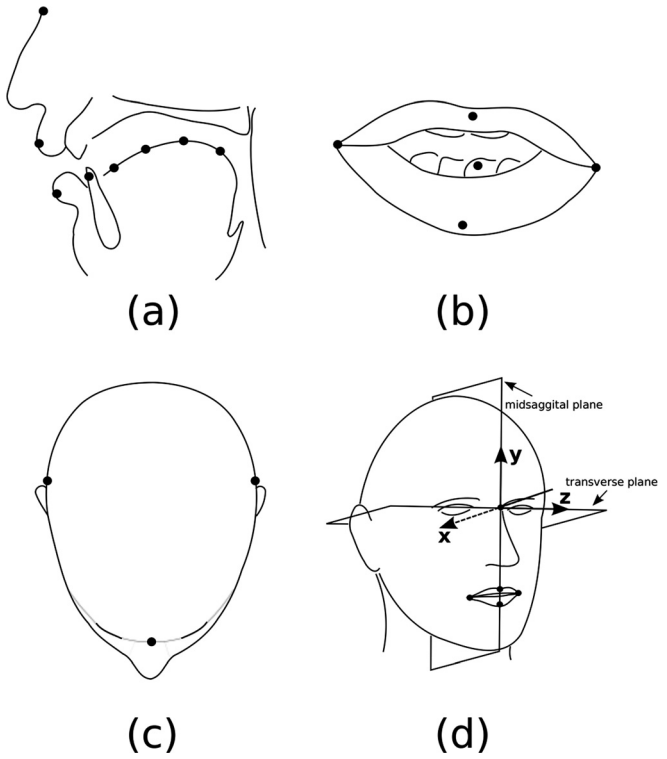

FIG. 2. EMA recording setup: (a) sagittal slice of the vocal tract, showing the sensors that were glued approximately on the midsagittal plane—on the tongue (four sensors), lower incisors, upper lip, lower lip, and bridge of the nose; (b) frontal view of the mouth showing sensors on the lip corners, upper lip, lower lip, and lower incisors; (c) view from the top of the head showing the three reference sensors on the bridge of the nose and behind the ears; (d) approximate sketch of the transverse and midsagittal planes, which are defined on the basis of the reference sensors' positions. The x (directed from the nose sensor to the midpoint between the two ear sensors) and y axes define the midsagittal plane. The y and z axes define the transverse plane.

Since the measurement grid system by Maeda is two-dimensional (2D) and lies on the midsagittal plane, the latter needs to be defined in terms of the 3D EMA measurement coordinate system so that sensor positions can be projected on it. After head movement compensation, the residual variances of the movement of three reference sensors are negligible. The positions of these sensors define a transverse plane. The line connecting the nose sensor with the midpoint between the two ear sensors is contained in this transverse plane and should also be contained in an assumed midsagittal plane. Therefore, we define the midsagittal plane as the one that is perpendicular to the transverse plane and contains the aforementioned line [see Fig. 2(d)].

We define a 3D orthogonal coordinate system with its origin at the nose sensor position; its x axis lies at the intersection of the two planes and is directed toward the midpoint between the ears; its y axis is contained in the midsagittal plane and faces upward; and its z axis is contained in the transverse plane. Finding the projection of any given sensor position on the midsagittal plane is equivalent to finding its projections on the x and y axes.

Okadome and Honda[24] have previously suggested that a sensor glued on the Adam's apple could monitor the height of the larynx. However, such a sensor would only record skin movement and would not be accurate for tracking the larynx height. With our recording setup, as described above, we do not have any information on the state of the larynx. Thus, we do not attempt to estimate the larynx height parameter of the model by Maeda in what follows.

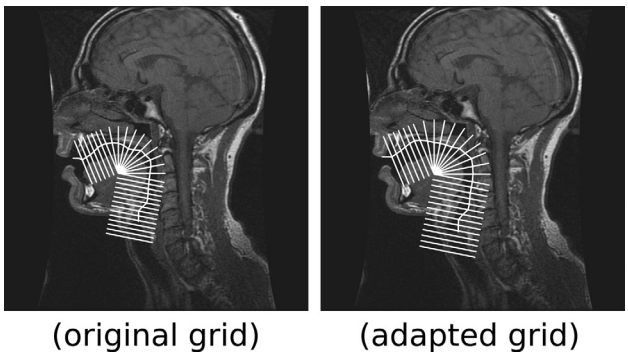

FIG. 3. Speaker adaptation of semipolar grid system. The adapted grid (right) derives from the original (left) after proper scaling and redrawing the external wall contour.

B. Model adaptation and data registration

The semipolar grid system and the external vocal tract provided by Maeda concern a specific female speaker. They had to be adapted to the male subject of this study for whom EMA data were recorded. The adaptation was performed manually assisted by a visualization software. As shown in Fig. 3, we superimposed the grid system on a midsagittal MRI image of the speaker and made two adjustments. First, we determined the mouth and pharynx scale factors and scaled the grid system accordingly. For our speaker the mouth scale was found to be 1.15 and the pharynx scale 1.20. Second, we corrected the external vocal tract contour by drawing our speaker's contour on the MRI image.

The coordinate system defined by the x and y axes derived in Sec. III A needs to be registered to the grid system by Maeda. To this end, we use EMA information on the shape of the palate. At the end of the recording session, a sensor was fixed onto a wooden stick and was used to trace the speaker's palate, approximately along the direction of the midsagittal plane.

We thus have two 2D descriptions of the palate. The first one is the just aforementioned EMA tracing, which is registered in the same coordinate system as the EMA data and projected on the previously derived midsagittal plane. The second one is the external vocal tract wall contour of the adapted articulatory grid. In order to register the EMA data to the grid, it is enough to register the first description of the palate to the second one.

We use an iterative pseudopoint matching algorithm[25] that repeats iteratively three steps. First, it determines a set of N minimal distance pairs between points of the two curves, denoted by $\{\mathbf{x}_i\}$ for the measured tracing and $\{\mathbf{y}_i\}$ for the model palate. Second, it finds a rotation matrix R and a shift vector t, so that the mean-squares objective function

$$\mathcal{F}(\mathbf{R},\mathbf{t}) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i||^2$$

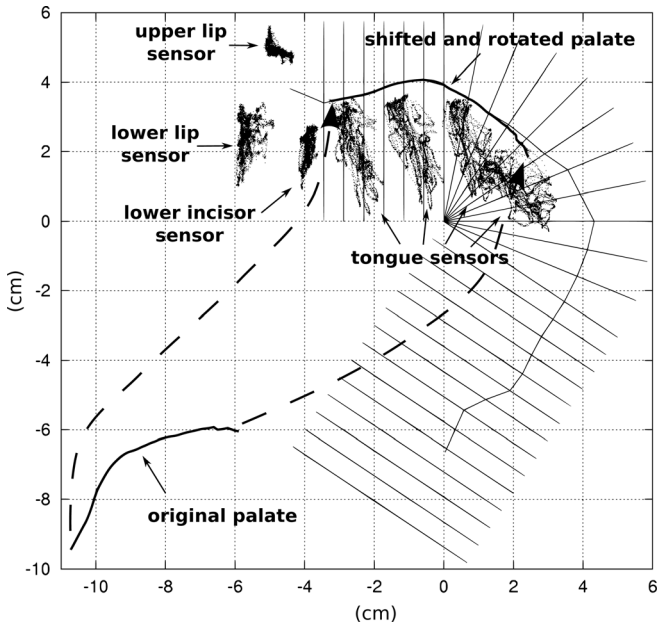

FIG. 4. Illustration of the registration process: The original EMA-traced palate was shifted and rotated, using an iterative algorithm, so that it matches the external wall of the model. The same shift and rotation was then applied to all EMA data. Registered data for upper and lower lip, lower incisor, and tongue sensors, corresponding to a part of the dataset including VV and VCV sequences, are shown as dots.

is minimized. Third, it applies the derived transformation on the coordinates of the points of the measured tracing. The iterations are repeated until there is no significant further rotation or shift. The total transformation is calculated as the concatenation of the transformations determined in every iteration.

In Fig. 4 we show the original (as projected on the midsagittal plane) and registered EMA trace of the palate with the semipolar grid system and the external vocal tract contour. We also superimpose on this drawing the registered EMA data, for the jaw, tongue, and upper and lower lip sensors corresponding to the full recording session previously described.

C. Extraction of articulatory variables

The tongue variables are normalized measurements of the tongue contour with respect to the gridlines of the semipolar coordinate system. EMA provides the positions of four sensors attached to the tongue, which are then registered to the grid system. We apply cubic spline interpolation[26] to the four tongue sensor positions and determine the intersections of the resulting curve with the gridlines between the front-most (tongue tip) and the back-most sensor, as well as with the gridlines just before the front-most sensor (toward the lips) and just after the back-most sensor (toward the tongue root, see Fig. 5). For normalization, we used means and standard deviations that were provided by Maeda and adapted to our speaker. This adaptation consisted of simple multiplication by the mouth scale factor (for the gridlines in the buccal linear area) or the pharynx scale factor (pharyngeal linear area) or a linear combination of the two (polar area).

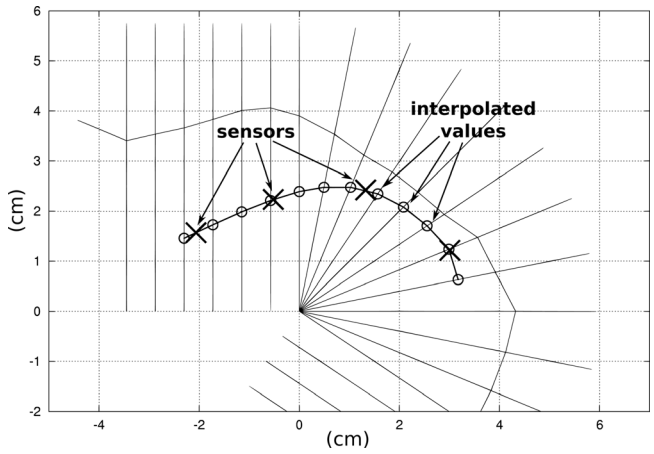

FIG. 5. Illustration of the procedure for getting the values of the tongue variables. Tongue sensor positions (crosses) in the semipolar grid space, derived spline (line) and its intersections (circles) with the gridlines.

An alternative process for the extraction of the tongue variables would involve the exploitation of the azimuth and elevation measurements for the tongue sensor, which could ideally provide the tangent of the tongue contour at each sensor position. However, this would require that the sensors are glued with their main axes exactly parallel to the midline of the tongue, which is quite difficult in practice.

Regarding the lip and jaw variables, we cannot readily use the definitions by Maeda (see Sec. II). Those definitions were based on x-ray data and supplementary photographs of the frontal view of the lips. They are not necessarily relevant in the EMA case. For example, Maeda used the front inner

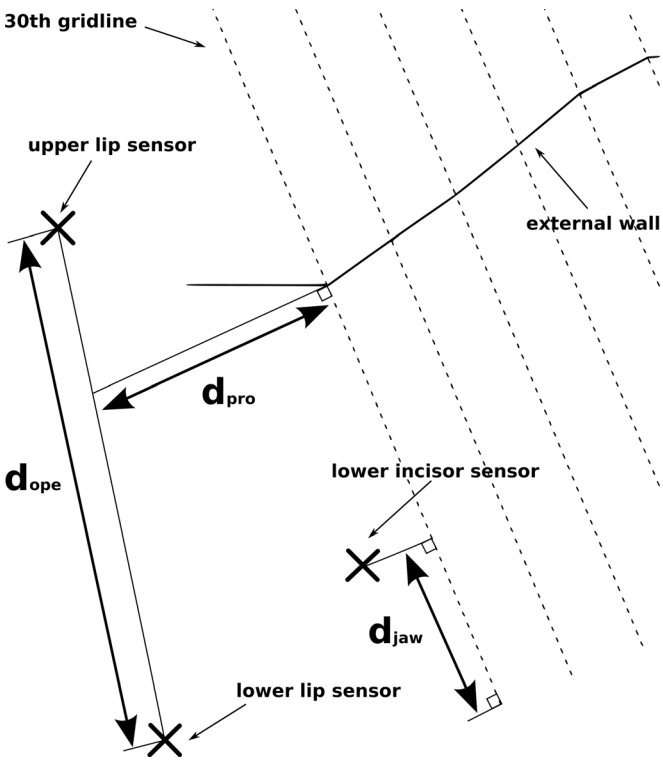

FIG. 6. Illustration of the definition of measurements d_{jaw}, d_{ope}, and d_{pro}.

lip contours to define lip opening. The exact lip flesh-points that contribute to the inner contour change over time. On the other hand, with EMA, we rely on the fixed flesh-points where the sensors were glued. Thus, the definitions by Maeda for the lip and jaw variables have to be adjusted in the EMA case.

As shown in Fig. 6, we take four measurements: (1) d_{jaw} is the position of the lower incisor sensor projected on the first gridline (the one closer to the lips); (2) d_{ope} is the distance between the upper and lower lip sensors; (3) d_{pro} is the distance along the direction that is perpendicular to the first gridline, between the intersection of the palatal contour with the first gridline and the line that connects the upper and lower lip sensor; (4) d_{wid} is the distance between the sensors on the lip corners.

Variables v_{jaw}, v_{pro}, and v_{wid} are normalized versions of measurements d_{jaw}, d_{pro}, and d_{wid}, respectively. The normalization factors (means and standard deviations) were determined as follows. For certain vowels, we know *a priori* the values that parameters P_1, P_5, and P_6 should take (shown in Table I). From these predefined values and matrix A_{lip}, we calculate the corresponding values for the jaw and lip variables. We take reference measurements of d_{jaw}, d_{pro}, and d_{wid} from the data and calculate their mean values. Then, we select normalization factors that fit these mean values to the values of Table I.

The lip opening variable does not exactly correspond to the d_{ope} measurement but is also influenced by d_{pro}. For example, in a transition from /i/ to /u/, the distance between the upper and lower lip sensors remains almost unaltered, while the lip opening variable (i.e., the effective lip opening) significantly decreases, as illustrated in Fig. 7.

To solve this problem we use the following strategy: First we take reference measurements of d_{ope} and d_{pro} from several instances of the phonemes /i/,/a/,/u/, and /p/. For the vowels, the target values for the normalized v_{ope} are shown in Table I. For the bilabial consonant, the raw value of v_{ope} should be equal to zero, which leads to a normalized value of -1.4. By fitting a linear model to these reference values we derive the empirical relationship

$$v_{\text{ope}} = 0.63 d_{\text{ope}} - 1.08 d_{\text{pro}} + 0.01.$$

We note that this expression depends on the exact positioning of the sensors.

D. Framewise recovery of parameters

Matrices A_{tng} and A_{lip} of Eq. (1) can be combined into a single matrix A so that

TABLE I. Suggested values of jaw (P_1) and lip (P_5, P_6) parameters for vowels /a/, /i/ and /u/ as provided by Maeda. Corresponding variable values are calculated using Eq. (1).

Vowel	P_1	P_5	P_6	v_{jaw}	v_{pro}	v_{ope}	v_{wid}
/a/	−1.5	0.5	−0.5	−1.5	−0.9	0.7	0.9
/i/	0.5	0.5	−1.0	0.5	−1.0	0.4	0.6
/u/	0.5	−1	1.5	0.5	1.6	−0.6	−1.0

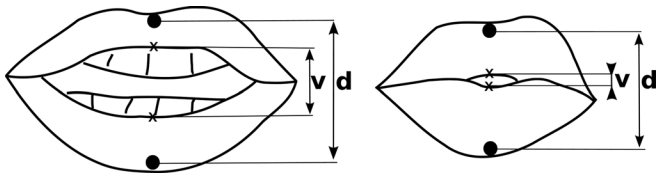

FIG. 7. Characteristic lip shapes for /i/ (left) and /u/ (right), taken from Bothorel *et al.* (Ref. 31) (pp. 19 and 77). Due to protrusion, the difference in effective lip opening (v) is much larger than what the raw measurement of the distance between the upper and lower lip sensors (approximately d) suggests.

$$\mathbf{v} = \mathbf{A}\mathbf{p}, \tag{2}$$

where

$$\mathbf{p} = [P_1, P_2, P_3, P_4, P_5, P_6]^T,$$

and

$$\mathbf{v} = \left[v_{\text{jaw}}, v_{\text{pro}}, v_{\text{ope}}, v_{\text{wid}}, v_{\text{tng6}}, v_{\text{tng7}}, \ldots, v_{\text{tng30}}\right]^T,$$

which we rewrite for convenience as

$$\mathbf{v} = [v_1, v_2, \ldots, v_{29}]^T.$$

Since the EMA sensors cover only a limited part of the tongue contour, only a limited number of tongue variables are available at any given instant. The exact positions of the corresponding elements in the v vector change over time. At a given instant, let $C \subset \{1, \ldots, 29\}$ be the positions in the v vector of the variables which are available.

According to Eq. (2) a vector of parameters p would generate the vector of variables V(p) with components V_i(p) such that

$$V_i(\mathbf{p}) = \sum_{j=1}^{6} a_{i,j} p_j,$$

where $a_{i,j}$ are elements of matrix A. Our objective is to find the set of parameters that generates variables V_i(p) with minimal distance to the observed ones v_i, i.e., we seek to minimize the quantity

$$I_{\text{s}} = \sum_{i \in C} \left(v_i - \sum_{j=1}^{6} a_{i,j} p_j \right)^2. \tag{3}$$

The minimization of I_{s} constitutes a typical least-squares problem. However, if solved unconstrained, it might give rise to parameter vectors that correspond to unrealistic vocal tract shapes. For plausibility of generated vocal tract configurations, the model requires that the z-scored parameters lie in the range [–3, 3]. Thus, the minimization of I_{s} should be subjected to the constraints

$$p_i \in [-3, 3],\ i = 1, \ldots, 6. \tag{4}$$

Additionally, we require that the produced tongue contours do not cross the motionless exterior tract wall. The corresponding constraint is

$$\mathbf{A}_{\text{tng}}\mathbf{p}_{\text{tng}} \leq \mathbf{w}, \tag{5}$$

where vector w describes the exterior wall.

Equation (3) can be easily expanded to the form

$$I_s = \mathbf{p}^T\mathbf{H}\mathbf{p} + 2\mathbf{p}^T\mathbf{q} + r,$$

where H is a symmetric matrix, q is a six-dimensional vector, and r is a constant. The minimization I_s subjected to the constraints of Eqs. (4) and (5) constitutes a quadratic programming problem that is readily solved using Octave's qp function[27] which implements an iterative active set null space method.[28]

E. Introduction of dynamic constraints

The method described so far ensures that the produced articulatory parameters satisfy two criteria. The first criterion is the proximity of the generated configurations to the EMA data, as expressed by Eq. (3). The second criterion is the generation of realistic vocal tract shapes, as imposed by the constraints of Eqs. (4) and (5). An additional criterion is included to ensure the smoothness of the produced articulatory trajectories dynamically. We apply a regularization method which uses the theory of variational calculus,[29] giving rise to an iterative process [30] that optimizes a cost function combining proximity to the measured variables and changing rate of articulatory parameters.

The parameters of the articulatory model are time functions $\mathbf{p}(t) = [P_1(1), P_2(t), \ldots, P_6(t)]^T, t \in [t_s, t_f]$ (remember that we have excluded the larynx height parameter of the model). The input data are time functions corresponding to the components of the vector v, denoted $v_i, i = 1, \ldots, 29$. Let $V_i(\mathrm{p}(t))$ be the components of the vector $\mathrm{V}(t) = \mathrm{A}\mathrm{p}(t)$, i.e., the variables generated by linear transformation of an assumed vector of articulatory parameters.

We introduce the following cost function which has to be minimized:

$$I_d = \int_{t_s}^{t_f} \sum_{i=1}^{29} h_i(t)[v_i(t) - V_i(\mathbf{p}(t))]^2 dt + \lambda \int_{t_s}^{t_f} \sum_{j=1}^{6} p\prime_j(t)^2 dt + \beta \int_{t_s}^{t_f} \sum_{j=1}^{6} p_j^2(t) dt\,, \tag{6}$$

where the function $h_i(t)$ accounts for the fact that not all variables are available at time t and is defined as follows:

$$h_i(t) = \begin{cases} 1, \text{if } v_j \text{ is available at time } t \\ 0, \text{otherwise.} \end{cases}$$

In Eq. (6), the first term expresses the proximity between observed variables $v_i(t)$ and those generated by the articulatory model $V_i(\mathrm{p}(t))$. The second term expresses the changing rate of articulatory parameters. The third is a potential energy term that penalizes large articulatory efforts and prevents the vocal tract from reaching positions too far from the equilibrium.

Following a procedure similar to the one we have detailed in previous work,[7] variational calculus is applied and leads to the equation:

$$\mathbf{B}\mathbf{p}_j^\tau = \mathbf{c}_j^\tau.$$

where B is the $(N+1) \times (N+1)$ matrix:

$$\mathbf{B} = \begin{bmatrix} \gamma+\beta+\lambda & -\lambda & 0 & \cdots & 0 \\ -\lambda & \gamma+\beta+2\lambda & -\lambda & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\lambda & \gamma+\beta+2\lambda & -\lambda \\ 0 & \cdots & 0 & -\lambda & \gamma+\beta+\lambda \end{bmatrix},$$

and the vectors $\mathbf{p}_j^\tau$ and $\mathbf{c}_j^\tau$ are defined as

$$\mathbf{p}_j^\tau = [p_{j,0}^\tau, \ldots, p_{j,k}^\tau, \ldots, p_{j,N}^\tau]^T,$$

and

$$\mathbf{c}_j^\tau = \begin{bmatrix} \gamma p_{j,0}^{\tau-1} - \sum_{i=1}^{29} h_i a_{i,j}(v_{i,0} - V_{i,0})\gamma p_{j,1}^{\tau-1} \\ -\sum_{i=1}^{29} h_j a_{i,j}(v_{i,1} - V_{i,1}) \\ \cdots \\ \gamma p_{j,N}^{\tau-1} - \sum_{i=1}^{29} h_j a_{i,j}(v_{i,N} - V_{i,N}) \end{bmatrix}.$$

This equation defines an iterative procedure that minimizes the expression of Eq. (6) At each iteration p_j is calculated for each of the six articulatory parameters. The startup solution is provided by the solution of the optimization problem described in Sec. III D.

IV. RESULTS

We applied our method to a large number of sequences across our recorded corpus, estimating articulatory control parameters from EMA data and generating the corresponding vocal tract shapes. Figure 8 presents examples of generated vocal tract shapes for French vowels /a/, /u/, /i/, and /e/, corresponding to instants close to the middle of the second vowel of V_1V_2 sequences pronounced at natural speed. The larynx parameter was fixed to zero. The derived shapes are shown superimposed on actual MRI images of the same speaker, uttering the same vowel, but sustained and in an isolated context. For the acquisition of these images the speaker articulated each vowel seven times. Each articulation lasted approximately 18 s, followed by a small pause so that the speaker could catch his breath. The speaker laid in a supine position during the acquisition.

There are two important criteria for evaluating the results of Fig. 8: (1) the correspondence between EMA

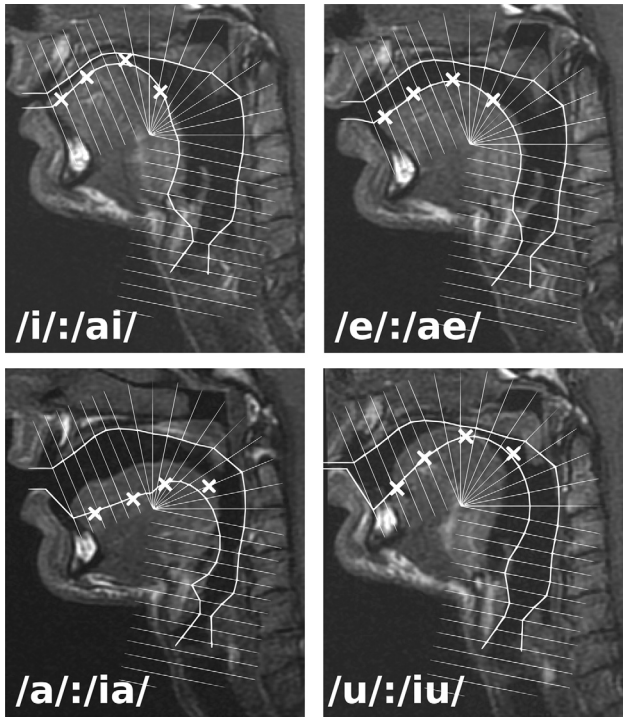

FIG. 8. EMA-derived vocal tract shapes for instances from V_2 of V_1V_2 sequences (dynamic utterances, upright position), superimposed on MRI images of the same vowels (sustained articulations, supine position). Crosses indicate the tongue sensor positions.

data presented by the four tongue sensors and the shape of the tongue obtained by the model; and (2) the quality of the shape of the back of the tongue obtained by the model. The first point might seem trivial, since the method is constructed to aim for closeness between the positions of the sensors and the shape of the tongue. However, there are cases where the derived tongue contour does not perfectly match the sensor position. This is the case, e.g., for the third sensor, counting from the lips, for /i/, or for the last sensor for /a/. A possible reason is that the sensors present a configuration which is beyond the representational capabilities of the model. That is, the model is unable to produce a shape that perfectly matches the positions of all sensors. The difficulty regarding the second point is that we do not have any EMA data regarding the shape of the back of the tongue, since it is extremely difficult to glue EMA sensors in the pharyngeal area. The question that naturally arises is whether the model predict a realistic shape of the back of the tongue. A positive answer to such a question would validate not only our method but also the reliability of the model by Maeda itself. Since the model was implemented based on a set of correlation coefficients obtained from repetitive procedures of targeted regressions for a certain dataset, some correlation between suprapharyngeal and pharyngeal regions was observed and modeled. Therefore, we should expect a certain degree of agreement in predicting the back of tongue. On the other hand, we should also expect some errors, since such correlation is not unique but averaged from many variations.

The articulations of /i/ and /e/ shown in Fig. 8, and their comparison to the corresponding MRI images, may be used to evaluate this point, since for these two cases the EMA data agree with the MRI images. Indeed, in these two cases, the shape of the back of the tongue predicted by the model is adequately similar to that presented by the MRI images.

However, the presented articulations of /a/ and /u/ would not offer a fair comparison since the shape of the front of the tongue provided by the EMA data is quite different to that of the MRI images. For /a/, these differences may be explained by arguing that the speaker, during MRI acquisition, probably reduced mouth opening to limit the output air flow in order to phonate the sound for 18 s. This is possible because /a/ presents high articulatory variability.[31] It also seems that the speaker chose a different articulatory strategy for the production of /u/ during the MRI and EMA acquisitions, probably due to the differences between the acquisition conditions, such as duration of articulation, context, supine versus upright position, and noise in the scanner. In previous work,[32] we had identified three possible articulatory strategies for /u/, based on inversion experiments: one with the narrowest constriction of the vocal tract in the palatal area, one with the constriction in the velar area, and one with the constriction in the uvular area[32] (see Fig. 9). The differences among formant frequencies for these shapes were well below 10 Hz. Though our MRI-presented /u/ is closer to the identified palatal one, our EMA-derived one is closer to the identified velar one. Moreover, the constriction of the /u/ presented by the MRI is at a more anterior position compared to tracings found in the literature, e.g., in Bothorel[31] or Delattre.[33]

We calculated area functions and then formant frequency values from the vocal tract slices of Fig. 8 using an articulatory-to-acoustic simulation proposed by Maeda.[34] This simulation works under a series of approximations regarding boundary losses due to friction and heat conduction, sagittal-to-area conversion, and nonrigidity of vocal tract walls.

Table II shows these model-derived formant frequencies. Additionally, it shows formant frequencies for the vowels /o/ and /y/, taken from the V_1V_2 sequences /io/ and /ay/. The larynx parameter was considered as fixed to zero. The table also shows the corresponding formant frequencies that were determined through observation of the spectrum from the real speech which was recorded concurrently with EMA. Finally, it shows generic vowel centers and standard deviations for male speakers of French, as given in the literature.[35]

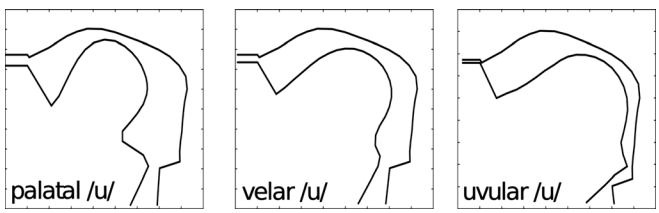

FIG. 9. Three possible vocal tract shapes for /u/, with approximately the same formant frequencies, as identified by inversion experiments (Ref. 32).

The vowel which exhibits the best results, in terms of consistency of the three sets of formants is /u/. The vowel with the worst results is /a/, especially for F1, where the value derived by the model is too low. Regarding /e/, the problem is that while the model-derived formants agree well with the speech-derived ones, there is no agreement with the generic vowel centers. Actually the speech-derived and model-derived formants are somewhere between the generic vowel centers for /e/ and /ɛ/. Given that the contrast between the two sounds is not very robust, even for native French speakers, it is possible that our speaker did not pronounce a clear /e/. For /i/,/o/, and /y/, there is a relatively good agreement between model-derived formants, speech-derived formants, and reference vowel centers since the differences between them are less than two standard deviations (as given in the table). Nonetheless, we should note that when speech was synthesized with a Klatt synthesizer[36] using these formant frequencies with constant amplitudes and bandwidths, and F0 copied from the concurrently recorded speech signal, the produced sounds were intelligible and identifiable, as indicated by informal listening tests.

Furthermore, we experimented with several values for the larynx parameter, keeping the rest of the parameters at the values derived from EMA. With our adapted model, the range from −3 to 3 for the value of the larynx parameter corresponds to a range of 2.7 cm in larynx height. Our findings were in accordance with the literature:[37,38] (1) Raising/lowering of the larynx results in raising/lowering of F1, F2, and F3; (2) The change in F2 for high front vowels is substantial; (3) The change in F3 for all studied vowels (oral French vowels) is also substantial.

A possible reason for some of the differences between model-derived and speech-derived formants might be some misestimation of the tongue-root region by the model. To further test this assumption, we did some simulation experiments on how sensitively the acoustics of the vowels presented in Fig. 8 would react to such a misestimation. Beginning with the reference shapes shown in Fig. 8 we changed the tongue contour uniformly between gridlines 8 and 17 [as shown in Fig. 1(b)], by 1 mm, 2 mm, or 1 cm, and measured the variation of the formant frequencies. More specifically, a change of 1 mm means that we subtracted the formant frequencies derived by narrowing the cavity by 0.5 mm from the formant frequencies derived by widening the cavity by 0.5 mm. The results of this experiment are shown in Table III.

TABLE III. Variation of formant frequencies with respect to a decrease of 1 mm, 2 mm, or 1 cm of the pharyngeal cavity width.

	1 mm			2 mm			1 cm		
Vowel	Δ F1	Δ F2	Δ F3	Δ F1	Δ F2	Δ F3	Δ F1	Δ F2	Δ F3
/i/	4	−29	−1	8	−57	−3	39	−285	−15
/e/	8	−25	−18	16	−50	−36	82	−261	−176
/a/	5	−21	−15	11	−43	−29	54	−220	−146
/u/	6	−2	−32	11	−3	−65	62	−29	−308

For all four studied vowels, as the pharyngeal cavity narrows, F1 increases. This increase is most relevant for /u/ and /i/, since F1 corresponds to the Helmholtz resonance of the back cavity. At the same time, F2 and F3 decrease. Their decrease is almost negligible for F2 of /u/ and F3 of /i/, and important in all other cases. These results also suggest a certain degree of linearity in the relationship between the first three formant values and the width of the pharyngeal cavity.

Can we use these results to confirm the assumption that the only cause of differences between model-derived and speech-derived trajectories in Table II is a possible misestimation of the contour at the tongue-root region? If we consider, for example, the case of /i/, we can see in Table III that a narrowing of the pharyngeal cavity (i.e., as if we had overestimated its width) would result to a desired increase in F1, a desired decrease in F2, and an undesired decrease in F3. Thus, the assumption cannot be confirmed and we should consider other possible sources of error in the model-derived formant frequencies, to complement or replace the tongue-root misestimation. The list of such sources of error includes the lack of information regarding the larynx, the epiglottis, and the velum, the lack of modeling of the sublingual cavity, the rough modeling of the lips, and the approximations under which the articulatory synthesizer operates.

In Fig. 10, we present an example of the dynamic aspects of articulation. The figure shows the estimated temporal evolution of the six model parameters for the sequence /iu/. We plot parameter trajectories obtained directly as a

TABLE II. The "model-derived" columns show the formant frequencies derived from the shapes of Fig. 8 via articulatory-to-acoustic simulation. The "speech-derived" columns show the tracked formant frequencies from the concurrently recorded speech. The last three columns show generic vowel centers and standard deviations for French male speakers as given in the literature (Ref. 35).

	Model-derived			Speech-derived			Vowel centers (deviations) (35)		
Vowel	F1	F2	F3	F1	F2	F3	F1	F2	F3
/i/	286	2092	2809	312	1930	3151	308 (34)	2064 (134)	2976 (147)
/e/	406	1729	2322	435	1737	2301	365 (31)	1961 (119)	2644 (107)
/a/	527	1198	2300	637	1260	2351	684 (47)	1256 (32)	2503 (131)
/u/	297	766	2107	302	781	2102	315 (43)	764 (59)	2027 (136)
/o/	345	904	2155	389	852	2152	383 (22)	793 (63)	2283 (126)
/y/	313	1618	2171	331	1686	2416	300 (37)	1750 (121)	2120 (182)

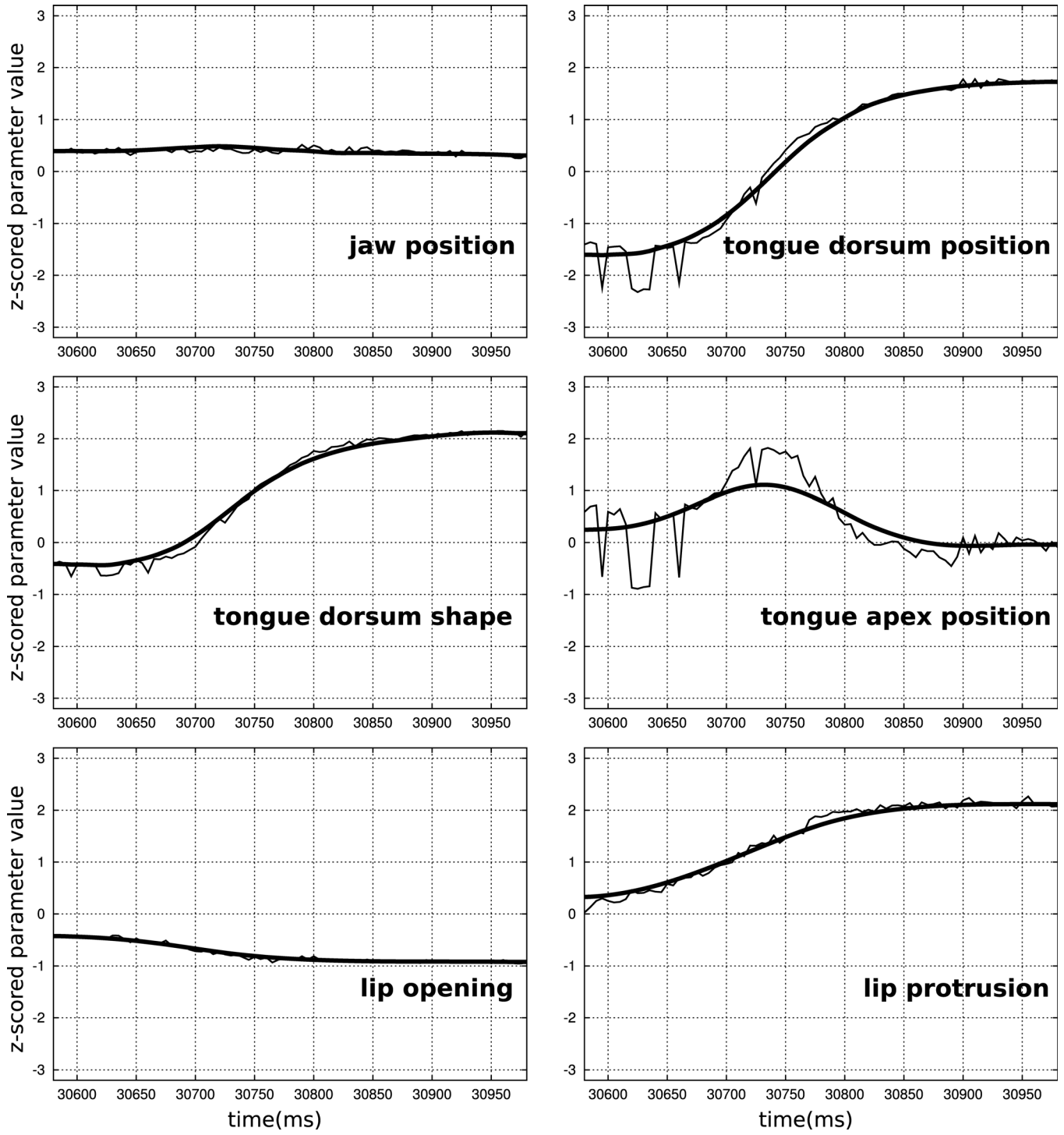

FIG. 10. Dynamic evolution of articulatory parameters for sequence /iu/. *z*-scored values are plotted against time in milliseconds. Thinner lines correspond to the solution before regularization, bolder lines to the regularized solution.

result of quadratic optimization and those after regularization. In fact, the proposed variational regularization method allows a trade-off between proximity to observed data and smoothness of the trajectories. It can have an important effect in cases where the quadratic optimization provides behavior that does not seem to be natural. For example, in the interval roughly between 30 550 and 30 680 ms both tongue dorsum position and tongue apex position parameter trajectories exhibit a series of discontinuities. In the interval between 30 680 and 30 800 ms the tongue apex position trajectory exhibits a bump that would indicate a nonsmooth transition between the two vowels. Applying regularization removes these outliers without affecting much the corresponding vocal tract shape sequence. This can be verified in Fig. 11, which presents the sequence of vocal tract shapes resulting from the parameter trajectories of Fig. 10. The outliers

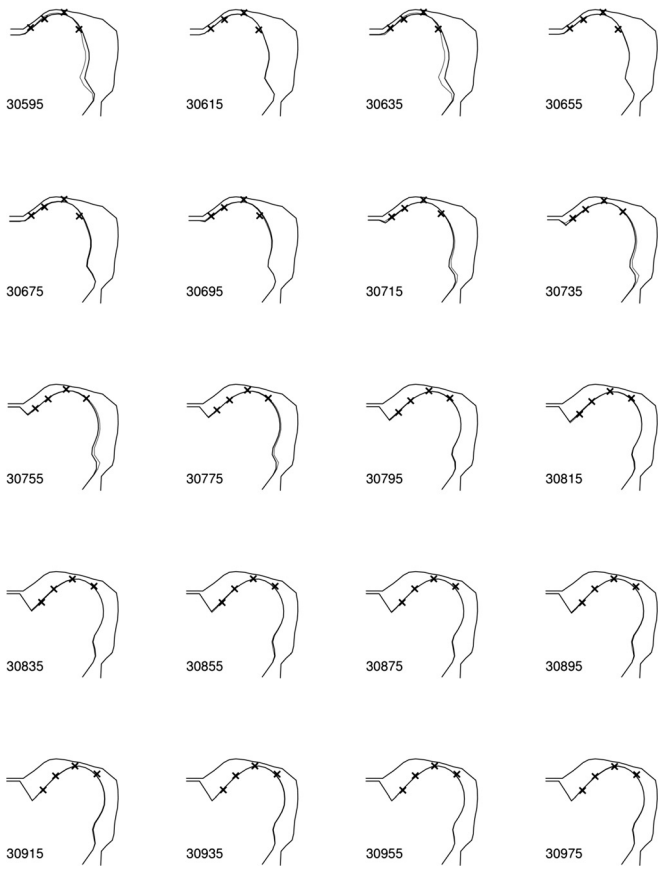

FIG. 11. Sequence of vocal tract shapes for sequence /iu/. Tongue sensor positions are marked. Time stamps (in milliseconds) are shown at lower left corner of each image. Thinner contours correspond to the solution before regularization, bolder contours to the regularized solution. For the sake of clarity, the temporal resolution for this figure is subsampled from 200 Hz (AG500 temporal resolution) to 40 Hz.

discussed previously and the effect of regularization can be observed by comparing the two plotted contours in frames labeled 30 595, 30 635, 30 735, and 30 755. Nevertheless, in most cases applying regularization does not lead to important

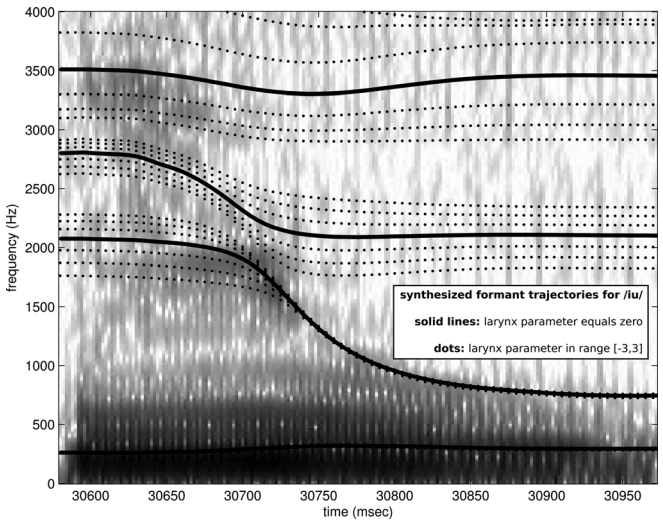

FIG. 12. Model-derived first four formant frequencies for sequence /iu/ superimposed on a spectrogram of the concurrently recorded speech. For solid lines, larynx height parameter is fixed at zero; for dots, it covers the range [−3,3].

modifications, except slightly smoothing the parameter trajectories, thus removing probable measurement noise.

Figure 12 shows the corresponding formant trajectories, when the larynx parameter is fixed to zero (solid lines), as well as when it covers the range [–3,3] (dots). Higher values for the larynx parameter lead to higher values for the formants. These trajectories are superimposed on a spectrogram of the concurrently recorded actual speech. It is known that rounded vowels such as /u/ have a lower larynx position than unrounded vowel such as /i/.[39] Applying this remark to Fig. 12, we may expect that for the /i/ part the actual formants lie at the upper part of the range, above the solid lines, and for /u/ at the lower part of the range.

V. CONCLUDING REMARKS

The articulatory representation provided by EMA is sparse, i.e., it concerns only a limited number of sensors attached to articulators. The control parameters of an articulatory model, such as the model by Maeda, constitute a more informative and easily interpretable articulatory representation compared to raw EMA information. Additionally, they allow direct speech synthesis. The method proposed in this paper may be regarded as an indirect way to map EMA information to speech acoustics: a problem that is otherwise addressed via data-driven statistical methods.[40,41]

Our main motivation for this work has been the evaluation of analysis-by-synthesis speech inversion methods.[7,32] Direct evaluation of such methods in the articulatory space is difficult, since relevant data are rare.[7] The method proposed in this work may create abundant information for the evaluation of such inversion methods. In addition it may allow the incorporation of constraints to our inversion framework. In previous work,[32] we introduced constraints on the space of articulatory solutions provided by the codebook method, based on standard phonetic knowledge. It is imaginable to replace those abstract constraints with new ones on the basis of EMA findings.

We have already mentioned that the effective lip opening is different than the measurement of the vertical distance between the upper and lower lip EMA sensors. We made the assumption that the effective lip opening is a linear combination of this vertical distance and the protrusion measurement. Perhaps this is an oversimplification: a more elaborate model might better account for the relationship between EMA measurements and the effective lip opening. This is an issue for further study.

The estimation of the larynx parameter is missing from our study. As previously said, we do not believe that an EMA sensor glued on the Adam's apple of the speaker would provide reliable information on the larynx height. A second modality, such as a video capture device, could be added to the recording setup for this purpose; however, elaborate registration algorithms would be required in such a case. Another possibility is to use several sensors glued in the neighborhood of Adam's apple to track the larynx height by reconstructing the skin surface in this area. In this case

the relative movement of the mouth cavity to the pharynx cavity (e.g., a nod gesture) should be removed.

The described method allows the recovery of midsagittal vocal tract shapes on the basis of the positions of a few fleshpoints. However, the diversity of these shapes is restricted by the model which simplifies vocal tract shape by suppressing fine articulatory details and operates under a series of assumptions that probably need further testing. Even though the model should generalize well to new speakers, we should expect some differences between the articulatory strategy of the speaker used in the construction of the model and that of the speaker used in our experiment. We must also bear in mind certain inaccuracies in the calculation of the positions of the sensors by the articulograph,[42,43] with a probable nonlinear effect on the estimated articulatory parameters. Finally, the approximations under which the articulatory-to-acoustic simulation works are not necessarily accurate. Thus, even if we accept a perfect estimation of the articulatory control parameters that correspond to the recorded data, we cannot expect a perfect match between the synthesized formant trajectories and those of the original speech. Improvements to the articulatory model, construction of new models, or alternative choices among the models existing in the literature, might alleviate some of these problems.

ACKNOWLEDGMENTS

We thank Shinji Maeda for fruitful discussions and for making his model and articulatory-to-acoustic simulation software available to us. We acknowledge the financial support of Contrat de Projets Etat-Région - Modélisation, Informations et Systémes Numérique for the acquisition of the electromagnetic articulograph.

[1]C. H. Coker, "A model of articulatory dynamics and control," Proc. IEEE **64**, 452–460 (1976).
[2]P. Mermelstein, "Articulatory model for the study of speech production," J. Acous. Soc. Am. **53**, 1070–1082 (1973).
[3]S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires (An articulatory model of the tongue with linear components)," in *Proceedings of Journées d'Étude de la Parole* (Days of Speech Study) (Grenoble, France, 1979), pp. 152–162.
[4]B. Gabioud, "Articulatory models in speech synthesis," in *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-of-the-Art and Future Challenges*, edited by E. Keller (Wiley, Chichester, United Kingdom, 1994), pp. 215–230.
[5]O. Engwall, "Modeling of the vocal tract in three dimensions," in *Eurospeech, Budapest* (1999), pp. 113–116.
[6]J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal," IEEE Trans. Speech Audio Process. **2**, 133–150 (1994).
[7]S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," J. Acous. Soc. Am. **118**, 444–460 (2005).
[8]S. Panchapagesan and A. Alwan, "Vocal tract inversion by cepstral analysis-by-synthesis using chain matrices," in *Interspeech, Brisbane* (2008), pp. 2857–2860
[9]B. Story, I. Titze, and E. Hoffman, "Vocal tract area functions from magnetic resonance imaging," J. Acous. Soc. Am. **100**, 537 (1996).
[10]S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acous. Soc. Am. **115**, 1771–1776 (2004).
[11]M. Stone, "Imaging the tongue and vocal tract," Int. J. Lang Commun. Disord. **26**, 11–23 (1991).
[12]M. Stone, "A guide to analysing tongue motion from ultrasound images," Clin. Linguist. Phonetics **19**, 455–501 (2005).
[13]M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, C. Kambhamettu, M. Li, V. Parthasarathy, and J. Prince, "Comparison of speech production in upright and supine position," J. Acous. Soc. Am. **122**, 532–541 (2007).
[14]P. Schonle, K. Grabe, P. Wenig, J. Hohne, J. Schrader, and B. Conrad, "Electromagnetic articulography—Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain Lang. **31**, 26–35 (1987).
[15]J. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta, and M. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," J. Acous. Soc. Am. **92**, 3078–3096 (1992).
[16]A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H. Tillmann, "Extracting tongues from moving heads," in *5th Speech Production Seminar*, Kloster Seeon, Germany (2000), pp. 313–316.
[17]T. Kaburagi, K. Wakamiya, and M. Honda, "Three-dimensional electromagnetic articulography—A measurement principle," J. Acous. Soc. Am. **118**, 428–443 (2005).
[18]T. Kaburagi and M. Honda, "Determination of sagittal tongue shape from the positions of points on the tongue surface," J. Acous. Soc. Am. **96**, 1356–1366 (1994).
[19]P. Badin, E. Baricchi, and A. Vilain, "Determining tongue articulation: from discrete fleshpoints to continuous shadow," in *Eurospeech*, Rhodes (1997), pp. 47–50.
[20]D. H. Whalen, A. M. Kang, H. S. Magen, R. K. Fulbright, and J. C. Gore, "Predicting midsagittal pharynx shape from tongue position during vowel production," J. Speech Lang. Hear. Res. **42**, 592–603 (1999).
[21]M.-T. Jackson and R. McGowan, "Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels—Statistical considerations," J. Acous. Soc. Am. **123**, 336–346 (2008).
[22]S. Maeda "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990), pp. 131–149.
[23]P. Combescure, "20 listes de dix phrases phonétiquement équilibrées (20 lists of ten phonetically balanced phrases)," Rev. Acoust. **56**, 34–38 (1981).
[24]T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," J. Acous. Soc. Am. **110**, 453–463 (2001).
[25]Z. Zhang, "Iterative point matching of free-form curves and surfaces," Int. J. Comput. Vis. **13**, 119–152 (1994).
[26]C. Ueberhuber, *Numerical Computation: Methods, Software, and Analysis* (Springer-Verlag, Berlin, 1997), pp. 412–416.
[27]J. Eaton, *GNU Octave: A High-Level Interactive Language for Numerical Computations* (Network Theory, Ltd., Bristol, United Kingdom, 1997), pp. 319–320.
[28]R. Fletcher, *Practical Methods of Optimization* (Wiley-Interscience, New York, 1987), pp. 240–245.
[29]M. Bonvalet, "Introduction au calcul des variations (Introduction to variational calculus)" in *Les principes variationnels (The Variational Principles)* (Masson, Paris, 1993), pp. 39–56.
[30]Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," in *ICASSP, Seattle* (1998), Vol. 2, pp. 929–932.
[31]A. Bothorel, P. Simon, F. Wioland, and J. Zerling, " Cinéradiographie des voyelles et consonnes du français (Cineradiographies of vowels and consonants in French)" (L'Institut de Phonétique de Strasbourg, France, 1986), pp. 38–45, 74–81.
[32]B. Potard, Y. Laprie, and S. Ouni, "Incorporation of phonetic constraints in acoustic-to-articulatory inversion," J. Acous. Soc. Am. **123**, 2310–2323 (2008).
[33]P. Delattre, "Pharyngeal features in the consonants of Arabic, German, Spanish, French, and American English," Phonetica **23**, 129–155 (1971).
[34]S. Maeda, "A digital simulation method of the vocal tract system," Speech Commun. **1**, 199–229 (1982).
[35]F. Lonchamp, "Description acoustique (Acoustic description)," in *La parole et son traitement automatic, par Calliope (Speech and Its Automatic Processing, by Calliope)*, edited by J. P. Tubach (Masson, Paris, 1989), pp. 79–130.
[36]D. Klatt, "Software for a cascade/parallel formant synthesizer," J. Acous. Soc. Am. **67**, 971–995 (1980).

[37]J. Sundberg and P.-E. Nordstrom, "Raised and lowered larynx-the effect on vowel formant frequencies," STL-QSPR **17**, 35–39 (1976).
[38]C. J. Riordan, "Control of vocal-tract length in speech," J. Acous. Soc. Am. **62**, 998–1002 (1977).
[39]P. Hoole and C. Kroos, "Control of larynx height in vowel production," in *ICSLP, Sydney* (1998), pp. 531–534.
[40]C. Kello and D. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," J. Acous. Soc. Am. **116**, 2354–2364 (2004).
[41]T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Commun. **50**, 215–227 (2008).
[42]C. Kroos, "Measurement accuracy in 3D electromagnetic articulography (Carstens AG500)," in *International Seminar on Speech Production*, Strasbourg, France (2008), pp. 61–64.
[43]Y. Yunusova, J. R. Green, and A. Mefferd, "Accuracy assessment for AG500, electromagnetic articulograph.," J. Speech Lang. Hear. Res. 547–555 (2009).